Queue-Aware Uplink Bandwidth Allocation and Rate Control for Polling Service in IEEE 802.16 Broadband Wireless Networks

Dusit Niyato, *Student Member*, *IEEE*, and Ekram Hossain, *Senior Member*, *IEEE*

Abstract—IEEE 802.16 standard defines the air interface specifications for broadband access in wireless metropolitan area networks. Although the medium access control signaling has been well-defined in the IEEE 802.16 specifications, resource management and scheduling, which are crucial components to guarantee quality of service performances, still remain as open issues. In this paper, we propose adaptive queue-aware uplink bandwidth allocation and rate control mechanisms in a subscriber station for *polling service* in IEEE 802.16 broadband wireless networks. While the bandwidth allocation mechanism adaptively allocates bandwidth for polling service in the presence of higher priority *unsolicited grant service*, the rate control mechanism dynamically limits the transmission rate for the connections under polling service. Both of these schemes exploit the queue status information to guarantee the desired quality of service (QoS) performance for polling service. We present a queuing analytical framework to analyze the proposed resource management model from which various performance measures for polling service in both steady and transient states can be obtained. We also analyze the performance of best-effort service in the presence of unsolicited grant service and polling service. The proposed analytical model would be useful for performance evaluation and engineering of radio resource management alternatives in a subscriber station so that the desired quality of service performances for polling service can be achieved. Analytical results are validated by simulations and typical numerical results are presented.

Index Terms—Broadband wireless networks, IEEE 802.16, dynamic bandwidth allocation, quality of service (QoS), queuing analysis.

1 Introduction

IEEE 802.16 standard-based broadband wireless access (BWA) technology is a promising alternative for last mile access in crowded urban areas or suburban areas where installation of cable-based infrastructure is economically or technically infeasible [1]. With advantages such as high transmission rate and predefined quality of service (QoS) framework, IEEE 802.16-based BWA is a viable technology to be used for connecting home networks and business LANs (e.g., IEEE 802.11) to the wired Internet. Also, this standard is evolving toward supporting nomadic and mobile users [2].

Radio resource management and scheduling mechanisms are crucial to guarantee QoS in IEEE 802.16 networks. However, these components were not specified in the standard. These mechanisms are important not only for the base station (BS), but also for a subscriber station (SS) to provide differentiated services among the different types of traffic corresponding to different connections. Since different types of traffic (e.g., constant-bit-rate, real-time, and best-effort) require different QoS performances, the radio resource management algorithm in an SS must allocate the available bandwidth among the different connections accordingly to meet the predefined QoS requirements.

In this paper, we propose queue-aware uplink bandwidth allocation and rate control schemes in an SS. These schemes can be applied for both real-time and non-real-time polling service (PS) as defined in the IEEE 802.16 specifications. Under the proposed bandwidth allocation scheme, the amount of bandwidth allocated for polling service can be adjusted dynamically according to the variations in traffic load and/or channel quality (and, hence, the queue length) so that the packet-level QoS performances such as protocol data unit (PDU) delay[1] and PDU dropping probability can be maintained at the desired level. Also, rate control is used to limit the transmission rate of the connections under polling service class so that the QoS performances can be controlled. The proposed queue-aware rate control scheme can be applied to each connection separately so that service differentiation (i.e., prioritization) among the connections can be achieved through different parameter settings.

A queuing analytical framework is presented to evaluate the performances of the proposed schemes. This is based on a discrete-time Markov chain which is formulated by considering Markov modulated Poisson process (MMPP) as the traffic sources under polling service. The advantages of using MMPP are two-fold: First, MMPP is able to capture the burstiness in the traffic arrival pattern, which is a common characteristic for multimedia and real-time traffic such as voice over IP (VoIP) and MPEG video [3] as well as Internet traffic [4]. Second, it is possible to obtain the MMPP parameters analytically for

• *The authors are with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada R3T 5V6. E-mail: {tao, ekram}@ee.umanitoba.ca.*

Manuscript received 20 Apr. 2005; revised 7 Sept. 2005; accepted 22 Oct. 2005; published online 17 Apr. 2006.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-0106-0405.

1. The IEEE 802.16 medium access control (MAC) uses a variable length protocol data unit (PDU) along with a number of other concepts that greatly increase the efficiency of data tranmission. Multiple MAC PDUs may be concatenated into a single burst to save physical layer (PHY) overhead.

1536-1233/06/$20.00 © 2006 IEEE Published by the IEEE CS, CASS, ComSoc, IES, & SPS

multiplexed traffic sources so that the queuing performances for multiple users can be analyzed.

The proposed radio resource management model for PS considers the impact of higher-priority traffic corresponding to unsolicited granted service (UGS) class for which the bandwidth can be statically or dynamically allocated according to the connections' transmission rate requirements. We also present an approximate queuing analytical model for best-effort (BE) service from which the basic performance measures (e.g., average delay) for BE traffic can be obtained as well as the impact of polling service on BE service can be investigated. We use simulations to validate the correctness of the analytical model.

The major contributions of this paper can be summarized as follows:

- A queue-state aware bandwidth allocation mechanism is proposed for reserving transmission bandwidth at a subscriber station for polling service. Also, a queue-state-based rate control method (both on aggregate and per-flow basis) is presented to limit the packet generation rate for connections under polling service.
- A queuing analytical model is developed to investigate the performances (under both steady state and transient state) of the queue-aware bandwidth allocation and the rate control mechanisms for polling service.
- An approximate queuing model is developed for analyzing the performance of best-effort traffic in presence of polling service.

2 Related Work

Radio resource scheduling (i.e., bandwidth allocation) and admission control are crucial for provisioning QoS in an 802.16 network. In [5], QoS-aware packet scheduling schemes were proposed for different types of traffic at the 802.16 base station. A resource allocation strategy, namely, enhanced staggered resource allocation (ESRA) method, was proposed in [6] with an objective of maximizing the number of concurrent transmissions so that the throughput can be maximized. However, the buffer dynamics at the radio link level queue (and, hence, the queuing performance) were not analyzed.

In [7], an admission control scheme for broadband multiservices wireless networks was presented to limit the number of ongoing connections so that the QoS for each connection can be maintained at the desired level. A dynamic resource allocation scheme for broadband orthogonal frequency division multiple access (OFDMA) networks was presented in [8], where the allocation is performed in two steps, namely, bandwidth allocation and channel assignment. Also, an M/G/1/K queuing model was used to estimate the packet blocking probability based upon which dynamic bandwidth allocation can be performed. The QoS differentiation was not considered in this work.

Although the general problem of radio resource management was studied extensively in the literature (e.g., in [9], [10], [11], [12]), the radio link level queuing aspects were ignored in most of the cases and the queuing dynamics (and, hence, the packet-level QoS) were not exploited for resource management and transmission rate control in wireless networks. The problem of optimal polling among several queues was studied in the literature. In [13], an optimal policy for polling (scheduling) was obtained to stochastically minimize the amount of unfinished work and the number of customers in the queues.

Rate control has been widely used in the wired-network environment to limit the transmission rate of the traffic sources. The performance of the rate control mechanism in ATM networks was studied in [14] by using a queuing model and the throughput degradation was quantified. Rate control can be implemented through the random early drop (RED) [15] mechanism to block the incoming packets gradually to avoid congestion. A proportional rate control mechanism for wireless networks was proposed in [15] to stabilize traffic oscillations. In [16], a theoretical model for wireless traffic control was proposed considering the impacts of congestion and error in the wireless channel. The model was developed based on rate-controlled earliest deadline first (RC-EDF) scheduling framework. However, these works did not consider multiple classes of connections with different QoS requirements.

The problem of analyzing radio link level queuing under wireless packet-transmission was addressed in the literature. In [18], a Markov-based model was presented to analyze the radio link level packet dropping process under ARQ-based error control. In [19], an analytical model to derive packet loss rate, average throughput, and average spectral efficiency under adaptive modulation and coding (AMC) was presented. However, all of these works considered only a *single-user* environment, which is significantly different from an IEEE 802.16-based system.

3 IEEE 802.16 Broadband Wireless Access Networks

3.1 Medium Access Control (MAC) and the Physical (PHY) Layers

In the IEEE 802.16 architecture [1], there are two types of stationary stations: subscriber station (SS) and base station (BS). The BS governs all communications to and from the subscriber stations. The physical layer of IEEE 802.16 operates in 10-66 GHz (IEEE 802.16) or 2-11 GHz (IEEE 802.16a) band and supports data rate in the range of 32-130 Mbps depending on the bandwidth of operation as well as the modulation and coding schemes. In the 10-66 GHz band, the signal propagation between BS and SS should be line-of-sight and single-carrier modulation is used. *WirelessMAN-SC* is the air interface specification for IEEE 802.16 operating in this frequency band. In contrast, IEEE 802.16a operates in the 2-11 GHz band and supports nonline-of-sight communication. The air interface specifications for 802.16a are: *WirelessMAN-SC2* for single-carrier modulation, WirelessMAN-OFDM for orthogonal frequency-division multiplexing (OFDM) with TDMA access scheme, and *WirelessMAN-OFDMA* for orthogonal OFDMA scheme.

In the 10-66 GHz band, channel bandwidth of 20, 25, or 28 MHz can be used. For modulation, QPSK, 16-QAM and 64-QAM can be used depending on the channel quality (i.e., signal-to-noise ratio (SNR) at the receiver). The system uses a frame size of 0.5, 1, or 2 ms for transmission and a frame is divided into subframes for downlink and uplink transmissions. While time-division multiplexing (TDM) is used for downlink transmission, time-division multiple access (TDMA) is used for uplink transmission. These subframes are composed of transmission bursts (i.e., uplink and

downlink bursts) which carry MAC information and users' data. Each transmission burst corresponding to a particular SS is separated from each other by a preamble field and contains several MAC PDUs. IEEE 802.16 uses a connection-oriented MAC which provides a mechanism for requesting bandwidth, transporting, and routing data to higher layer. IEEE 802.16 MAC supports two classes of SS: grant per connection (GPC) and grant per SS (GPSS). In the case of GPC, bandwidth is granted to a connection individually. In contrast, for GPSS, a portion of the available bandwidth is granted to each of the SSs and each SS is responsible for allocating the bandwidth among the corresponding connections. Since the BS does not need to keep track of allocations for all connections, GPSS is more scalable and efficient than GPC.

The lengths of uplink and downlink subframes are determined dynamically by the BS and are broadcast to the SSs through downlink and uplink map messages (UL-MAP and DL-MAP) at the beginning of each frame. Therefore, each SS knows when and how long to receive from and transmit data to BS. In the uplink direction, each SS can request bandwidth to BS by using BW-request PDU. There are two modes to transmit BW-request PDUs: contention mode and contention-free (polling) mode. In the contention mode, an SS transmits a BW-request PDU during predefined contention period and a backoff mechanism is used to resolve contention among the BW-request PDUs from multiple SSs. In contrast, in the contention-free mode, the BS polls each SS. After receiving the polling message, SSs respond by sending BW-request PDU. Due to predictable delay, the contention-free mode is suitable for QoS sensitive applications.

Although IEEE 802.16 specifications define signaling for the multiple access mechanism, resource management and scheduling remain as open issues.

3.2 Service Types in IEEE 802.16

IEEE 802.16 defines the following three major types of services each of which has different QoS requirements.

3.2.1 Unsolicited Grant Service (UGS)

This type of service supports constant-bit-rate (CBR) traffic and multimedia traffic (e.g., VoIP). For this service type, the BS generally allocates a fixed amount of bandwidth to each of the connections in a static manner.

3.2.2 Polling Service (PS)

This service supports traffic for which QoS guarantee is required and it can be divided into two subtypes: real-time and non-real-time. The difference between these subtypes lies in the tightness of the QoS requirements (i.e., real-time PS is more sensitive to delay than non-real-time PS). Not only delay sensitive traffic but also non-real-time Internet traffic can use PS to achieve a certain throughput guarantee. The amount of bandwidth required for this type of service is determined dynamically based on the required QoS performances and the traffic arrival rates for the corresponding connections.

3.2.3 Best-Effort Service (BE)

This is for traffic with no QoS guarantee. The amount of bandwidth allocated for BE service depends on the bandwidth allocation policies for the other two types of service. In particular, the bandwith left after serving UGS and PS traffic is allocated for BE service.

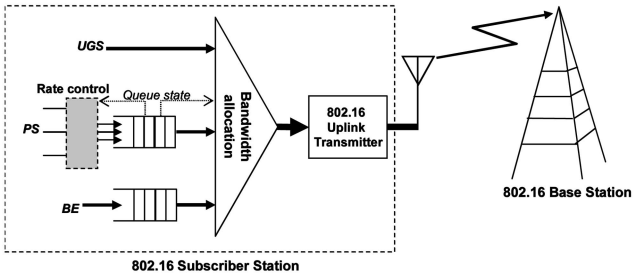


Fig. 1. System model.

4 System Model and Assumptions

4.1 System Description

We consider an uplink transmission scenario from an SS to a BS through the time-division multiple access (TDMA)/ time-division duplex (TDD) access mode and single carrier modulation (e.g., as in WirelessMAN-SC) for three traffic types, namely, UGS, PS, and BE traffic (Fig. 1). For PS, a dedicated queue is used to buffer the PDUs from the corresponding connections. We consider an SS of type GPSS for which a certain amount of bandwidth is reserved by the BS. This allocated bandwidth is shared among the different service types in the same SS, with UGS having the highest priority and the BE service having the lowest priority.

For better scalability, the PDUs from all the PS connections are aggregated into a single queue of size X PDUs. For the PS queue, rate control can be applied to control traffic at the packet-level. If the rate control parameters for each of the connections are identical, all PS connections experience the same QoS performance. Since there is no performance guarantee for best-effort traffic, the queue size for the best-effort traffic is assumed to be infinity.

4.2 Queue-Aware Bandwidth Allocation

We denote by b_{max} ($b_{max} \in \mathbb{N}$) the maximum number of MAC PDUs that an SS can transmit per uplink transmission subframe. We consider two modes of bandwidth allocation for PS, namely, complete partitioning (CP) and complete sharing (CS). With complete partitioning, a fixed amount of bandwidth b_{ugs} (from the total bandwidth allocated to an SS) is statically allocated for UGS while the remaining bandwidth (i.e., $b_{max} - b_{ugs}$) is allocated for PS and BE service. In case of complete sharing, when the bandwidth requirement for UGS traffic is less than b_{ugs}, the remaining available bandwidth will be available for PS. After the required amount of bandwidth has been allocated to UGS and PS traffic, the left over bandwidth is allocated to BE traffic.

We propose an uplink bandwidth allocation scheme for PS which takes the current number of PDUs in the PS queue into account. The allocation is done on a frame by frame basis in which the amount of bandwidth is determined for each transmission frame individually. In this scheme, the set of thresholds for bandwidth allocation is defined as follows:

$$\Psi = \{\psi_1, \psi_2, \cdots, \psi_b, \cdots, \psi_{b_{max}-b_{ugs}}\}, \quad (1)$$

where $\psi_b \in \{1, \cdots, X\}$, $\psi_b < \psi_{b+1}$, and $b \in \{1, \cdots, b_{max} - b_{ugs}\}$. This set of thresholds is used to indicate the amount of bandwidth required in each uplink subframe. In particular, the amount of bandwidth allocated to polling service is calculated as a function of number of PDUs in the PS queue for

complete partitioning and complete sharing schemes, respectively, as follows:

$$\mathcal{B}_{CP}(x) = \begin{cases} 0, & x = 0 \\ b, & \psi_b \le x < \psi_{b+1} \\ b_{max} - b_{ugs}, & \psi_{b_{max}-b_{ugs}} \le x \end{cases}$$
$$\mathcal{B}_{CS}(x) = \begin{cases} 0, & x = 0 \\ b, & \psi_b \le x < \psi_{b+1} \\ b_{max}, & \psi_{b_{max}-b_{ugs}} \le x. \end{cases} \quad (2)$$

4.3 Queue-Aware Rate Control

We propose a queue-aware rate control mechanism for PS connections in which the PDU arrival rate is controlled according to the number of PDUs in the queue. This rate control can be implemented either at the traffic source or at the PS queue. In the former case, the SS informs the traffic sources of the queue status.[2] In the latter case, rate control can be implemented similar to the random early detection (RED) mechanism [15] in an Internet router in which some PDUs received at the PS queue are randomly dropped.

Let $\tau_{min}, \tau_{max} \in \mathbb{N}$ denote the rate control thresholds for the number of PDUs in the queue and λ_{min} denote the minimum guaranteed arrival rate. Specifically, the transmission rate of traffic source under PS cannot be reduced below λ_{min}. Then, with a PDU arrival rate of λ_o, the rate control policy can be expressed as a function of the number of PDUs in the PS queue (x) as follows:

$$\tilde{\lambda}(x, \lambda_o, \lambda_{min}) = \begin{cases} \lambda_o, & x < \tau_{min} \\ \mathcal{F}(\lambda_o, x), & \tau_{min} \le x < \tau_{max} \\ \lambda_{min}, & \tau_{max} \le x, \end{cases} \quad (3)$$

where $\tilde{\lambda}(.)$ is the controlled arrival rate and $\mathcal{F}(\lambda, x)$ is a nonincreasing function of x with the constraint $\lambda_{min} \le \mathcal{F}(\lambda, x) \le \lambda_o$.

The rate control mechanism can be applied either on aggregate or on per-flow basis. In the former case, PDU arrival rates for all connections under PS are controlled using the same values of τ_{min}, τ_{max}, and λ_{min}. In the latter case, different parameter settings for rate control are used for each connection (i.e., $\tau_{min}(i)$, $\tau_{max}(i)$, and $\lambda_{min}(i)$ for connection i). While per-flow rate control is able to differentiate the QoS among different connections, aggregate rate control is simpler to implement and applicable when all connections have the same QoS requirements.

4.4 Error Control

To ensure the reliability of PDU transmission from SS to BS, an infinite persistent ARQ-based error recovery is used. That is, the erroneous PDUs will be retransmitted until they are successfully received at the BS. If θ denotes the PDU error rate (PER), assuming an independent error process, the probability that n PDUs out of m transmitted PDUs are successfully received can be obtained as follows:

$$\theta_{n,m} = \binom{m}{n}(1-\theta)^n(\theta)^{m-n}, \quad n \in \{0, 1, \cdots, m\}. \quad (4)$$

2. Note that, since the SS and the traffic sources are in the same local network, we ignore the delay incurred for updating queue status.

We also assume that the transmission status for the PDUs transmitted in the previous frame time is made available to the transmitter before transmissions in the current frame time start.

5 Queuing Analytical Model for Polling Service (PS)

We present an analytical framework based on a discrete-time Markov chain (DTMC) for the above bandwidth allocation and rate control schemes to analyze the QoS performance measures for the PS queue taking ARQ-based error control also into account. First, we describe the traffic models for UGS and PS traffic. Then, the queuing model for PS is formulated for both the cases of complete partitioning and complete sharing. The queuing model is solved to obtain the steady state and the transient state probabilities from which the QoS measures are obtained.

5.1 PDU Arrival Process for PS Connections

We assume that the PDU arrival process for each PS connection follows an MMPP model. An MMPP model is more general than a traditional Poisson model and is able to capture burstiness in the traffic arrival process. With MMPP, the PDU arrival rate λ_s is determined by the state s of the Markov chain and the total number of states is S (i.e., $s = 1, 2, \cdots, S$). The MMPP process for connection i can be represented by $\mathbf{U}(i)$ and $\mathbf{\Lambda}(i)$, in which the former is the transition probability matrix of the modulating Markov chain and the latter is the matrix of Poisson arrival rate. These matrices are defined as follows:

$$\mathbf{U}(i) = \begin{bmatrix} u_{1,1} & \cdots & u_{1,S} \\ \cdots & \cdots & \cdots \\ u_{S,1} & \cdots & u_{S,S} \end{bmatrix}, \mathbf{\Lambda}(i) = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_S \end{bmatrix}. \quad (5)$$

Discrete-time MMPP (dMMPP) [20] is equivalent to MMPP in the continuous time. In this case, the rate matrix $\mathbf{\Lambda}(i)$ is represented by diagonal probability matrix $\mathbf{\Lambda}_a(i)$ when the number of PDUs arriving in one frame is a. Note that $a \in \{0, 1, \cdots, A\}$, in which A is the maximum batch size for PDU arrival (e.g., $1 \le A \le \infty$). Each diagonal element of $\mathbf{\Lambda}_a(i)$ can be obtained from

$$f_a(\lambda_s) = \frac{e^{-\lambda_s T}(\lambda_s T)^a}{a!}, \quad (6)$$

where $f_a(\lambda_s)$ is the probability that a Poisson events occur during time interval T (i.e., frame length) with mean rate λ_s.

In the case of aggregated traffic from two users (e.g., user 1 and user 2), the matrices corresponding to state transition and PDU arrival probability for this multiplexed source can be calculated as follows:

$$\mathbf{U} = \mathbf{U}(1) \otimes \mathbf{U}(2), \quad (7)$$
$$\mathbf{\Lambda}_a = \sum_{i+j=a} \mathbf{\Lambda}_i(1) \otimes \mathbf{\Lambda}_j(2), \quad i, j \in \{0, 1, \cdots, A\}, \quad (8)$$

for $a = 0, 1, \cdots, A \times 2$, where $\otimes$ denotes Kronecker product. For the case with more than two users, these two matrices

can be obtained in a similar way. The average PDU arrival rate for connection i is obtained as follows:

$$\rho(i) = \boldsymbol{\pi}_m(i)\left(\sum_{a=0}^{A} a\boldsymbol{\Lambda}_a(i)\right)\mathbf{1}, \tag{9}$$

where $\boldsymbol{\pi}_m(i)$ is obtained by solving $\boldsymbol{\pi}_m(i)\mathbf{U}(i) = \boldsymbol{\pi}_m(i)$ and $\boldsymbol{\pi}_m(i)\mathbf{1} = 1$. Note that $\mathbf{1}$ is a column matrix of ones. Therefore, with a total of N connections, the total average PDU arrival rate at the PS queue can be obtained as follows:

$$\rho = \boldsymbol{\pi}_m\left(\sum_{a=0}^{N\times A} a\boldsymbol{\Lambda}_a\right)\mathbf{1} \tag{10}$$

and $\boldsymbol{\pi}_m$ is obtained from $\boldsymbol{\pi}_m\mathbf{U} = \boldsymbol{\pi}_m$ and $\boldsymbol{\pi}_m\mathbf{1} = 1$.

5.2 PDU Arrival Process for UGS Connections

For modeling the PDU arrival process for UGS connections, we consider a multistate on-off model which is a special type of dMMPP. The maximum number of states for each connection is C and the number of PDU arrivals when the source is in state c is c. While the state transition matrix $\mathbf{V}(i)$ of the multistate on-off model for connection i is similar to that of MMPP, the PDU arrival probability matrices $\boldsymbol{\Gamma}_c(i)$ are different. In particular, the maximum batch size is C (i.e., $A = C$) and the diagonal elements of these matrices are defined as follows:

$$[\boldsymbol{\Gamma}_c(i)]_{j,j} = \begin{cases} 1, & j = c+1 \\ 0, & \text{otherwise,} \end{cases} \qquad c \in \{0, 1, \cdots, C\}, \tag{11}$$

where the first row corresponds to the case of no PDU arrival and $[\boldsymbol{\Gamma}_c(i)]_{j,k}$ denotes the element at row j and column k of matrix $\boldsymbol{\Gamma}_c(i)$. If there are multiple UGS connections, the state transition matrix $\mathbf{V}$ and the PDU arrival probability matrices $\boldsymbol{\Gamma}_c$ of the multiplexed connection can be obtained from (7) and (8). Note that b_{ugs} denotes the maximum total bandwidth for UGS, where $b_{ugs} = M \times C$ for a total of M multistate on-off sources.

5.3 Formulation of the Queuing Model for Polling Service

In our queuing model, the state of the PS queue (i.e., the number of PDUs in the polling service queue) is observed at the beginning of each frame time. A PDU arriving during frame time f will not be transmitted until the next frame time $f+1$ at the earliest. The state space of the queue can be defined as follows:

$$\Delta = \{(\mathcal{S}, \mathcal{C}, \mathcal{X}),\ , 1 \le \mathcal{S} \le N \times S, 1 \le \mathcal{C} \le M, 0 \le \mathcal{X} \le X\}, \tag{12}$$

where $\mathcal{S}$ denotes the state of dMMPP traffic sources, $\mathcal{C}$ denotes the state of multistate on-off sources, and $\mathcal{X}$ is the number of PDUs in the PS queue. While the states of dMMPP and multistate on-off models are independent for all connections, the number of PDUs in the queue depends on the dMMPP arrival probabilities, the bandwidth usage for UGS connections, and the service rate at the PS queue (and, hence, the amount of allocated bandwidth). Also, the amount of allocated bandwidth depends on the number of PDUs in the PS queue and the set of thresholds Ψ. Note that, in case of complete partitioning, the model does not need to maintain the state of any multistate on-off source, and therefore, $\mathcal{C} = \{\emptyset\}$.

The transition matrix $\mathbf{P}$ of the queue can be expressed as in (13), where the rows of matrix $\mathbf{P}$ correspond to the number of PDUs in the PS queue (i.e., $\mathcal{X}$).

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}_{0,0} & \cdots & \mathbf{p}_{0,A\times N} & & \\ \vdots & \vdots & \ddots & \ddots & \\ \mathbf{p}_{b_{max},0} & \cdots & \cdots & \mathbf{p}_{b_{max},b_{max}+A\times N} & \\ & \ddots & \ddots & \ddots & \ddots \\ & \mathbf{p}_{y,y-b_{max}} & \cdots & \cdots & \mathbf{p}_{y,y+A\times N} \\ & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{13}$$

Matrices $\mathbf{p}_{x,x'}$ represent the changes in the number of PDUs in the queue (i.e., there are x PDUs during the current frame time and it will be x' during the next frame time).

5.3.1 Arrival Process under Rate Control

With queue-aware rate control, the matrix for the Poisson arrival process $\boldsymbol{\Lambda}(i)$ in the MMPP model for connection i depends on the number of PDUs in the PS queue. Therefore, we can express this matrix as follows:

$$\boldsymbol{\Lambda}^{(x)}(i) = \begin{bmatrix} \tilde{\lambda}(x, \lambda_1, \lambda_{min}) & & \\ & \ddots & \\ & & \tilde{\lambda}(x, \lambda_S, \lambda_{min}) \end{bmatrix}. \tag{14}$$

Then, the matrix $\boldsymbol{\Lambda}_a^{(x)}(i)$ is obtained by using (6). If there are multiple traffic sources, (7) and (8) are used to obtain the complete PDU arrival process (i.e., $\mathbf{U}$ and $\boldsymbol{\Lambda}_a^{(x)}$) at the PS queue. Note that (14) can be used for both aggregate and per-flow-based rate control.

5.3.2 Transition Matrix for the Complete Partitioning (CP) Model

In the case of complete partitioning, PDU arrival probability and dMMPP state transitions are given by $\mathbf{U}$ and $\boldsymbol{\Lambda}_a^{(x)}$. However, the PDU departure probabilities corresponding to all arrival state of dMMPP are identical and depend only on the number of PDUs in the queue and the PDU transmission error rate. Therefore, the probability of departure of n PDUs ($n \in \{0, 1, \cdots, b_{max} - b_{ugs}\}$) when there are x PDUs ($x \in \{0, 1, \cdots, X\}$) in the queue is obtained as follows:

$$\left[\mathbf{D}_n^{(x)}\right]_{j,j} = \theta_{n,\mathcal{B}_{CP}(x)}, \tag{15}$$

where $j \in \{1, 2, \cdots, S \times N\}$. Note that every matrix $\mathbf{D}_n^{(x)}$ has the same size as that of $\mathbf{U}$. Each element $\mathbf{p}_{x,x'}$ of matrix $\mathbf{P}$ in case of complete partitioning is obtained as follows:

$$\mathbf{p}_{x,x-g} = \mathbf{U} \sum_{\{n,a|n-a=g\}} \left(\boldsymbol{\Lambda}_a^{(x)} \times \mathbf{D}_n^{(x)}\right), \tag{16}$$

$$\mathbf{p}_{x,x+h} = \mathbf{U} \sum_{\{n,a|a-n=h\}} \left(\boldsymbol{\Lambda}_a^{(x)} \times \mathbf{D}_n^{(x)}\right), \tag{17}$$

$$\mathbf{p}_{x,x} = \mathbf{U} \sum_{\{n,a|n=a\}} \left(\boldsymbol{\Lambda}_a^{(x)} \times \mathbf{D}_n^{(x)}\right), \tag{18}$$

for $g = 1, 2, \cdots, G$ and $h = 1, 2, \cdots, A \times N$, where $n \in \{0, 1, 2, \cdots, G\}$ and $a \in \{0, 1, 2, \cdots, A \times N\}$ represent the number of departed PDUs and the number of PDU arrivals, respectively.

Considering both the arrival and the departure events, (16), (17), and (18) above represent the transition probability matrices for the cases when the number of PDUs in the queue decreases by g, increases by h, and does not change, respectively. Since the maximum total allocated bandwidth can be greater than the number of PDUs in the queue and the decrease in the number of PDUs cannot be less than the number of PDUs in the queue, the maximum amount by which the number of PDUs in the queue can decrease is obtained from $G = \min\left(b_{max} - b_{ugs}, x\right)$.

5.3.3 Transition Matrix for the Complete Sharing (CS) Model

In this case, we have to consider transmission of multistate on-off traffic for UGS connections which have higher priority and affect the bandwidth allocation for the PS traffic. The departure probability matrix for the multistate on-off sources (corresponding to the UGS connections) can be established as follows:

$$\left[\mathbf{E}_n^{(x)}\right]_{c+1,c+1} = \begin{cases} \theta_{n,m}, & m = \min\left(\mathcal{B}_{CS}(x), b_{max} - c\right) \\ 0, & \text{otherwise,} \end{cases} \quad (19)$$

where $c \in \{0, 1, \cdots, b_{ugs}\}$. Note that every matrix $\mathbf{E}_n^{(x)}$ has the same size as that of $\mathbf{V}$. For the CS case, each element $\mathbf{p}_{x,x'}$ of matrix $\mathbf{P}$ is obtained as follows:

$$\mathbf{p}_{x,x-g} = \mathbf{U} \otimes \mathbf{V} \sum_{\{n,a|n-a=g\}} \left(\mathbf{\Lambda}_a^{(x)} \otimes \mathbf{E}_n^{(x)}\right), \quad (20)$$

$$\mathbf{p}_{x,x+h} = \mathbf{U} \otimes \mathbf{V} \sum_{\{n,a|a-n=h\}} \left(\mathbf{\Lambda}_a^{(x)} \otimes \mathbf{E}_n^{(x)}\right), \quad (21)$$

$$\mathbf{p}_{x,x} = \mathbf{U} \otimes \mathbf{V} \sum_{\{n,a|n=a\}} \left(\mathbf{\Lambda}_a^{(x)} \otimes \mathbf{E}_n^{(x)}\right), \quad (22)$$

where $n \in \{0, 1, 2, \cdots, G\}$, $a \in \{0, 1, 2, \cdots, A \times N\}$, and $G = \min\left(b_{max}, x\right)$.

5.3.4 PDU Blocking Process

If the PS queue does not have enough space to accommodate all of the incoming PDUs, some of the PDUs will be blocked. In this case, the bottom part (i.e., the rows corresponding to the condition $(A \times N) + x > X$) of matrix $\mathbf{P}$ has to capture the PDU blocking effect. Therefore, (17) and (21), which correspond to the CP and the CS cases, respectively, become

$$\mathbf{p}_{x,x+h} = \sum_{i=h}^{A \times N} \hat{\mathbf{p}}_{x,x+i} \quad \text{for } x + h \geq X \quad (23)$$

and, for $x = X$, (18) and (22) become

$$\mathbf{p}_{x,x} = \hat{\mathbf{p}}_{x,x} + \sum_{i=1}^{A \times N} \hat{\mathbf{p}}_{x,x+i} \quad \text{for } x = X, \quad (24)$$

where $\hat{\mathbf{p}}_{x,x}$ is obtained for the case without PDU dropping.

5.3.5 Steady State Probabilities

The queuing performance measures for the PS traffic can be obtained from the steady state probability matrix $\boldsymbol{\pi}_{st}$ which is obtained by solving the equations

$$\boldsymbol{\pi}_{st}\mathbf{P} = \boldsymbol{\pi}_{st}, \quad \boldsymbol{\pi}_{st}\mathbf{1} = 1, \quad (25)$$

where $\mathbf{1}$ is a column matrix of ones. The matrix $\boldsymbol{\pi}_{st}$ contains steady state probabilities for the feasible combinations of the state variables $\mathcal{S}$, $\mathcal{C}$, and $\mathcal{X}$. This matrix can be decomposed into $\pi_{st}^{(CS)}(s, c, x)$, which is the steady state probability that the aggregated source is in state s, the multistate on-off source is in state c and there are x PDUs in the PS queue. Note that $\boldsymbol{\pi}_{st}$ is a row matrix and $[\boldsymbol{\pi}_{st}]_i$ indicates the element at column i of matrix $\boldsymbol{\pi}_{st}$. Since, in the case of complete sharing, the system state does not keep track of multistate on-off sources, the steady state probability is reduced to $\pi_{st}^{(CP)}(s, x)$.

5.3.6 Transient State Probabilities

In this section, we investigate the system behavior in the transient state. A system exhibits transient behavior when it is not in the steady state, i.e., during the transition period until the system reaches an equilibrium state [17]. Transient analysis is important to observe the system behavior with changes in inputs or system parameters, especially in a time-varying system which may rarely reach the steady state. Based on *Chapman-Kolmogorov* equations, the probability matrix of system states during frame time f can be obtained from

$$\boldsymbol{\pi}_{tr}(f) = \boldsymbol{\pi}_{tr}(f-1)\mathbf{P}(f), \quad (26)$$

where $\mathbf{P}(f)$ is the transition matrix during frame time f. The transient state probabilities $\pi_{tr}^{(CS)}(s, c, x, f)$ and $\pi_{tr}^{(CP)}(s, x, f)$ can be obtained in the same way as that for the steady state probabilities.

5.4 QoS Measures for Polling Service

Since the QoS measures for PS in both steady and transient states can be obtained in a similar way, we use $\pi^{(CS)}(s, c, x)$ and $\pi^{(CP)}(s, x)$ for the complete sharing and the complete partitioning cases, respectively, to represent the general probability that the dMMPP is in state s, the on-off source is in state c, and there are x PDUs in the PS queue.

5.4.1 Average Queue Length

The average queue length for the CP and the CS cases can be obtained as follows:

$$\overline{x}^{(CP)} = \sum_{x=0}^{X} x \left(\sum_{s=1}^{S \times N} \pi^{(CP)}(s, x)\right), \quad (27)$$

$$\overline{x}^{(CS)} = \sum_{x=0}^{X} x \left(\sum_{s=1}^{S \times N} \sum_{c=1}^{(b_{ugs}+1)} \pi^{(CS)}(s, c, x)\right). \quad (28)$$

5.4.2 Average PDU Arrival Rate

For the CP and the CS cases, this can be calculated for connection i as follows:

$$\overline{\rho}^{(CP)}(i) = \sum_{x=0}^{X} \left(\boldsymbol{\pi}_m(i) \left(\sum_{a=0}^{A} a\mathbf{\Lambda}_a^{(x)}(i)\right) \mathbf{1}\right) \sum_{s=1}^{S \times N} \pi^{(CP)}(s, x), \quad (29)$$

$$\overline{\rho}^{(CS)}(i) = \sum_{x=0}^{X} \left(\boldsymbol{\pi}_m(i) \left(\sum_{a=0}^{A} a\mathbf{\Lambda}_a^{(x)}(i)\right)\mathbf{1}\right) \sum_{s=1}^{S \times N} \sum_{c=1}^{(b_{ugs}+1)} \pi^{(CS)}(s, c, x). \quad (30)$$

The total average PDU arrival rate at the PS queue for the CP and the CS cases are calculated from

$$\overline{\rho}^{(CP)} = \sum_{i=1}^{N} \overline{\rho}(i)^{(CP)}, \quad \overline{\rho}^{(CP)} = \sum_{i=1}^{N} \overline{\rho}(i)^{(CP)}. \quad (31)$$

5.4.3 PDU Blocking Probability

To obtain the PDU blocking probability, we first calculate the average number of blocked PDUs per frame time [18]. Given that there are x PDUs in the PS queue and the queue size increases by h, if $h + x > X$, the number of blocked PDUs during one frame time is $h - (X - x)$ and zero otherwise. The average number of blocked PDUs per frame time for the complete partitioning and the complete sharing cases are obtained as follows:

$$\overline{x}_{bl}^{(CP)} = \sum_{s=1}^{S\times N}\sum_{x=0}^{X}\sum_{h=X-x+1}^{S\times N-\mathcal{B}_{CP}(x)} \pi^{(CP)}(s,x)\left(\sum_{j=1}^{S\times N}\left[\mathbf{p}_{x,x+h}\right]_{s,j}\right)(h-(X-x)), \quad (32)$$

$$\overline{x}_{bl}^{(CS)} = \sum_{s=1}^{S\times N}\sum_{c=1}^{(b_{ugs}+1)}\sum_{x=0}^{X}\sum_{h=X-x+1}^{S\times N-\mathcal{B}_{CS}(x)} \pi^{(CS)}(s,c,x)\left(\sum_{j=1}^{S\times N+(b_{ugs}+1)}\left[\mathbf{p}_{x,x+h}\right]_{s,j}\right)(h-(X-x)). \quad (33)$$

The terms $(\sum_{j=1}^{S\times N}[\mathbf{p}_{x,x+h}]_{s,j})$ in (32) and $(\sum_{j=1}^{S\times N+(b_{ugs}+1)}[\mathbf{p}_{x,x+h}]_{s,j})$ in (33) indicate the total probability that the number of PDUs in the queue increases by h at every state of the dMMPP and the multistate on-off sources. This probability differs from the probability of dMMPP arrival because we have to consider the number of PDUs transmitted during the same frame time as well. After calculating the average number of blocked PDUs per frame time, we can obtain the probability that an incoming PDU is blocked for the CP and the CS cases, respectively, as follows:

$$P_{bl}^{(CP)} = \frac{\overline{x}_{bl}^{(CP)}}{\overline{\rho}^{(CP)}}, \quad P_{bl}^{(CS)} = \frac{\overline{x}_{bl}^{(CS)}}{\overline{\rho}^{(CS)}}. \quad (34)$$

5.4.4 Queue Throughput

This gives the average number of transmitted PDUs per frame time. We calculate the throughput by using the fact that, if a PDU is not blocked upon its arrival, it will be transmitted eventually. Hence, the queue throughput (number of PDUs/frame interval) for the complete partitioning and the complete sharing cases can be obtained from

$$\eta^{(CP)} = \overline{\rho}^{(CP)}(1 - P_{bl}^{(CP)}), \quad \eta^{(CS)} = \overline{\rho}^{(CS)}(1 - P_{bl}^{(CS)}). \quad (35)$$

5.4.5 Average Allocated Bandwidth

For the proposed adaptive queue-aware bandwidth allocation, the average bandwidth allocated for the complete partitioning and the complete sharing cases can be obtained from

$$\overline{b}^{(CP)} = \sum_{x=0}^{X}(\mathcal{B}_{CP}(x))\left(\sum_{s=1}^{S\times N}\pi^{(CP)}(s,x)\right), \quad (36)$$

$$\overline{b}^{(CS)} = \sum_{x=0}^{X}\sum_{c=1}^{(b_{ugs}+1)}\left(\sum_{s=1}^{S\times N}(\mathcal{B}_{CS}(x)-(c-1))\pi^{(CS)}(s,c,x)\right). \quad (37)$$

5.4.6 Bandwidth Utilization

This performance measure indicates the utilization of allocated bandwidth and can be obtained from

$$\mu^{(CP)} = \frac{\eta^{(CP)}}{\overline{b}^{(CP)}}, \quad \mu^{(CS)} = \frac{\eta^{(CS)}}{\overline{b}^{(CS)}}. \quad (38)$$

5.4.7 Delay Statistics

Delay for a PDU is defined as the time interval (in terms of frames) since the PDU arrived at the queue and the time that it has been successfully transmitted. By applying Little's law, average delay is obtained from

$$\overline{d}^{(CP)} = \frac{\overline{x}^{(CP)}}{\eta^{(CP)}}, \quad \overline{d}^{(CS)} = \frac{\overline{x}^{(CS)}}{\eta^{(CS)}}. \quad (39)$$

6 Queuing Model for Best-Effort (BE) Service

In this section, we present a model for approximate analysis of the basic performance measures (e.g., average queuing delay) for best-effort service, which has the least priority among the three service classes. Since the allocated bandwidth for the BE queue depends on the state of the UGS and PS connections and the number of PDUs in the PS queue, the state space for the BE queue can be expressed as follows:

$$\Delta_{BE} = \{(\mathcal{S},\mathcal{C},\mathcal{X},\mathcal{Y}),\ 0 \le \mathcal{X} \le X, \mathcal{Y} \ge 0\}, \quad (40)$$

where $\mathcal{Y}$ is the number of PDUs in the BE queue with infinite buffer size. However, maintaining all these states will make the model quite complicated. Therefore, we present an approximate model with the simplified state space for the BE queue as follows:

$$\Delta_{BE} = \{(\mathcal{Y}),\ \mathcal{Y} \ge 0\}, i > 0. \quad (41)$$

The model is approximate in the sense that the correlation among multistate on-off sources, dMMPP sources, and number of PDUs in the PS queue is ignored. However, we will show later in this paper that the basic performance measures obtained from this approximate model are close to those obtained from simulations. The presented model is for the complete partitioning case. However, the model for complete sharing can be developed in a similar way.

We assume that the PDU arrival process is Poisson with average rate λ_{BE}. The maximum bandwidth that can be allocated to the BE queue is denoted by $B = b_{max} - b_{ugs}$. The transition matrix $\mathbf{Q}$ for this model can be obtained as in (42).

$$\mathbf{Q} = \begin{bmatrix} q_{0,0} & \cdots & q_{0,A} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ q_{B,0} & \cdots & \cdots & q_{B,B+A} & & \\ & \ddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \quad (42)$$

Note that, since this matrix $\mathbf{Q}$ is used to represent the number of packets in the BE queue, which is infinite, the structure of $\mathbf{Q}$ is different from $\mathbf{P}$ in (13).

Element $q_{y,y'}$ indicates the probability that the BE queue has y PDUs during the current frame time and it becomes y' in

the next frame time. To obtain this probability, we calculate the probability of departure of a PDU from the BE queue based on the number of PDUs in the PS queue as follows:

$$k_n = \sum_{x=\psi_b}^{\psi_{b+1}-1} \left(\sum_{s=1}^{S\times N} \pi^{(CP)}(s,x) \right), \tag{43}$$

for $n = B - b$, $b \in \{0, 1, \cdots, B\}$ and zero otherwise. Then, each element $q_{y,y'}$ is obtained as follows:

$$q_{y,y-g} = \sum_{\{n,a|n-a=g\}} f_a(\lambda_{BE}) \times k_n, \tag{44}$$

$$q_{y,y+h} = \sum_{\{n,a|a-n=h\}} f_a(\lambda_{BE}) \times k_n, \tag{45}$$

$$q_{y,y} = \sum_{\{n,a|n=a\}} f_a(\lambda_{BE}) \times k_n, \tag{46}$$

for $g = 1, 2, \cdots, G$ and $h = 1, 2, \cdots, A$, where $G = \min(B, y)$. Note that (44), (45), and (46) represent the transition probability matrices for the cases when the number of PDUs in the queue decreases by g, increases by h, and does not change, respectively.

Since the size of matrix $\mathbf{Q}$ is infinite, we apply the *matrix-geometric method* [21] to obtain the steady state probabilities. For this, we reblock matrix $\mathbf{Q}$ to obtain the transition probability matrix in the following form:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{K} & \mathbf{L} & & & \\ \mathbf{M} & \mathbf{N}_1 & \mathbf{N}_0 & & \\ & \mathbf{N}_2 & \mathbf{N}_1 & \mathbf{N}_0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{47}$$

When the stability condition, namely, $\boldsymbol{\delta}\mathbf{N}_2\mathbf{1} > \boldsymbol{\delta}\mathbf{N}_0\mathbf{1}$, where $\boldsymbol{\delta} = \boldsymbol{\delta}\mathbf{N}$, $\boldsymbol{\delta}\mathbf{1} = 1$, and $\mathbf{N} = \mathbf{N}_0 + \mathbf{N}_1 + \mathbf{N}_2$ is satisfied, then the matrix $\mathbf{R}$, which is the minimal nonnegative solution of $\mathbf{R} = \mathbf{N}_0 + \mathbf{R}\mathbf{N}_1 + \mathbf{R}^2\mathbf{N}_2$, can be determined such that $\boldsymbol{\zeta}_{i+1} = \boldsymbol{\zeta}_i\mathbf{R}$, where $\boldsymbol{\zeta}_i$ contains steady state probabilities corresponding to the number of PDUs in the BE queue. This matrix $\mathbf{R}$ can be obtained iteratively from

$$\mathbf{R}(k+1) = \mathbf{N}_0 + \mathbf{R}(k)\mathbf{N}_1 + \mathbf{R}^2(k)\mathbf{N}_2 \tag{48}$$

until $|\mathbf{R}(k+1) - \mathbf{R}(k)|_{i,j} < \epsilon$, $\forall i, j$ (e.g., $\epsilon = 10^{-9}$). Next, we calculate $\boldsymbol{\zeta}_0$ and $\boldsymbol{\zeta}_1$ by solving the following equations:

$$\mathbf{B}[\mathbf{R}] = \begin{bmatrix} \mathbf{K} & \mathbf{L} \\ \mathbf{M} & \mathbf{N}_1 + \mathbf{R}\mathbf{N}_2 \end{bmatrix}, \quad [\boldsymbol{\zeta}_0, \boldsymbol{\zeta}_1] = [\boldsymbol{\zeta}_0, \boldsymbol{\zeta}_1]\mathbf{B}[\mathbf{R}], \tag{49}$$

$$\boldsymbol{\zeta}_0\mathbf{1} + \boldsymbol{\zeta}_1(\mathbf{I} - \mathbf{R})^{-1}\mathbf{1} = 1. \tag{50}$$

Since $\boldsymbol{\zeta}_i$ consists of $A - 1$ states of different number of PDUs in the BE queue, the steady state probability of y PDUs in the BE queue $\zeta(y)$ can be extracted as follows:

$$\zeta(y) = [\boldsymbol{\zeta}_i]_{col(i,y)}, \text{ where } col(i,y) = i \times (A-1) + y + 1. \tag{51}$$

In this case, the calculation needs to be truncated at Y_t PDUs such that $1 - \sum_{y=0}^{Y_t} \zeta(y) < \epsilon$.

Then, the average number of PDUs in the BE queue $\overline{y}_{BE}$ and the average delay $\overline{d}_{BE}$ for a PDU in the BE queue can be simply obtained from

$$\overline{y}_{BE} = \sum_{y=0}^{Y_t} y\zeta(y), \quad \overline{d}_{BE} = \frac{\overline{y}_{BE}}{\lambda_{BE}}. \tag{52}$$

7 Performance Evaluation

7.1 Parameter Setting

We consider a TDMA/TDD-based uplink transmission scenario from a particular SS to the BS. The SS under consideration is stationary and works in GPSS mode. The communication between SS and BS uses *rate ID 0* [22] (i.e., QPSK with code rate 1/2). The PDU arrival process for each PS connection is assumed to be identical and it follows a two-state MMPP model (i.e., $S = 2$) with the following parameters:

$$\mathbf{U}(i) = \begin{bmatrix} 0.1 & 0.9 \\ 0.2 & 0.8 \end{bmatrix}, \boldsymbol{\Lambda}(i) = \alpha \begin{bmatrix} 1 & 0 \\ 0 & 2.2 \end{bmatrix}, \quad i = 1, \cdots, N, \tag{53}$$

where α indicates the traffic intensity and the maximum batch size of PDU arrival is 20 (i.e., $A = 20$). When $\alpha = 1$, the average PDU arrival rate of this dMMPP connection is $\rho(i) = 1.9818$. The PDUs from all PS connections are aggregated into the PS queue and the size of this queue is assumed to be 100 PDUs (i.e., $X = 100$). The transmitter serves the PS queue in a first-in-first-out fashion.

In our performance evaluation, we use $\alpha = 1.5$, bandwidth allocated to SS is 12 units (i.e., $b_{max} = 12$), the number of connections under PS is 2 (i.e., $N = 2$), and probability of successful transmission of a PDU is 0.995 (i.e., $\theta = 0.005$). For UGS traffic, we use a three-state on-off source with the transition matrix defined as follows:

$$\mathbf{V} = \begin{bmatrix} 0.3 & 0.7 & 0.0 \\ 0.2 & 0.5 & 0.3 \\ 0.0 & 0.5 & 0.5 \end{bmatrix}. \tag{54}$$

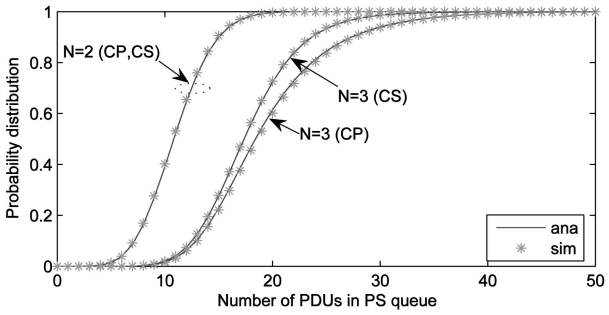
Therefore, this source requires bandwidth of 1.1667 units on average and b_{ugs} is set to 2. Note that we vary some of these parameters according to the evaluation scenarios, while the rest remain fixed according to the aforementioned setting.

For queue-aware bandwidth allocation, we consider sets of thresholds which are uniformly located over the range of buffer size. We use the notation $e_1 : (e)$ for the set of thresholds $\Psi = \{e_1, e_1 + e, e_1 + 2e, \cdots, e_1 + (b_{max} - b_{ugs})e\}$. For example, $1 : (e = 5)$ represents the set $\Psi = \{1, 6, 11, 16, 21, 26\}$, where $(b_{max} - b_{ugs}) = 6$. For queue-aware rate control, we assume that the minimum guaranteed rate is a function of the original rate and is defined as follows: $\lambda_{min} = \lambda_o/2$. Also, we assume $\mathcal{F}(\lambda_o, x) = \lambda_o \times \frac{\lambda_o(x - \tau_{min})}{2(\tau_{max} - \tau_{min})}$.

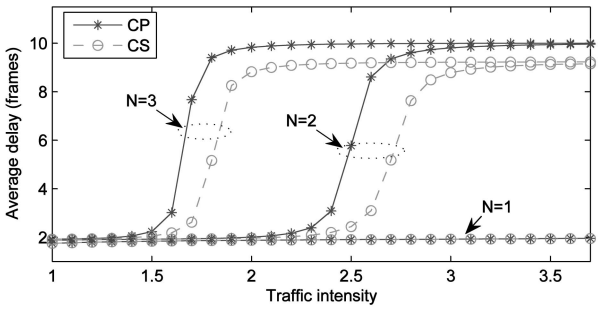
7.2 Simulation Environment

A time-driven simulator, developed in *MATLAB*, is used to evaluate the performance of the proposed queue-aware uplink bandwidth allocation and rate control algorithms and also to validate the correctness of the analytical model.

The PDU arrival and transmission events occur in a time-slotted fashion in which the length of a time-slot is equal to one frame interval. In the simulator, information on the states of a multistate on-off source (i.e., dMMPP for each connection) is maintained separately for each connection. The



(a)



(b)

Fig. 2. (a) Queue-length distribution and (b) average delay for the PS queue.

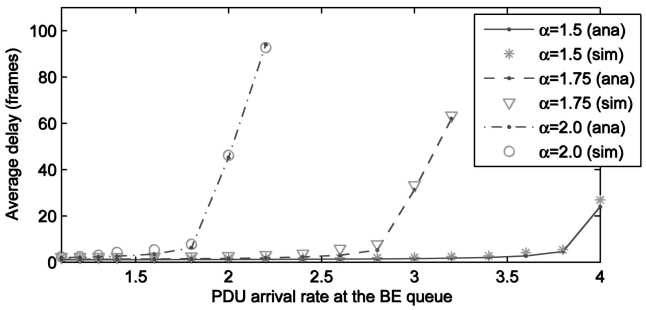


Fig. 3. Average delay for the BE queue.

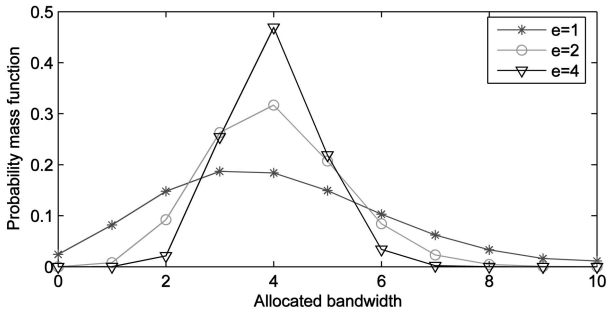


Fig. 4. Probability mass function for allocated bandwidth under different threshold settings.

number PDUs in the PS queue is calculated by considering the number of incoming PDUs for every time slot according to the state of dMMPP sources. In one time slot, the amount of bandwidth allocated to UGS service is determined based on the state of multistate on-off source. The remaining bandwidth is allocated to PS. Then, according to the threshold for bandwidth allocation setting (i.e., Ψ), the bandwidth left from PS will be allotted to BE service.

For queue-aware rate control of traffic arriving into the PS queue, some of the arriving PDUs are randomly blocked so that the arrival rate for the PS connections conforms to the desired setting (i.e., $\tilde{\lambda}(x, \lambda_o, \lambda_{min})$).

An independent packet error process is simulated for each wireless transmission. We replicate each simulation 10 times and, for each replication, the length of the simulation time is 200,000 time slots. We obtain the performance results for the bandwidth allocation (i.e., CP and CS schemes) and rate control scheme under varying traffic intensity, different settings of the bandwidth adaptation thresholds, and different rate control thresholds. Also, the performances of the proposed queue-aware bandwidth allocation schemes are compared with those of static allocation.

7.3 Numerical and Simulation Results

7.3.1 Queue-Length Distribution and Average Delay

Typical results on queue-length distributions and average queuing delay for both the CS and the CP cases are shown in Fig. 2. As expected, the length of the PS queue grows with the number of connections. Also, since the PS traffic can use the unused bandwidth from UGS traffic (e.g., when the multistate on-off source is in the off state), with the same number of PS connections, the queue length for the CS scheme is smaller than that for the CP case. However, for a small number of PS connections (e.g., $N = 2$), the queue-length distributions become very close to each other (Fig. 2a) since the transmission rate is high enough to accommodate all arriving PDUs. We observe that the simulation results follow the analytical results very closely which confirms the correctness of the analytical model.

The average delay increases with the number of PS connections (Fig. 2b). The average delay of the CS scheme is better than that of the CP scheme since, with CS, the bandwidth which is not used by UGS will be yielded to polling service. We observe that, when the traffic intensity is low and the number of PS connections is small, average delay remains constant since the transmission rate is high enough so that the delay remains constant over a range of values of traffic intensity.

However, when the traffic intensity reaches a certain point, which we call *critical rate*, average delay increases rapidly to the maximum delay. This steep rise occurs when the queue status changes from stable to unstable since the PDU arrival rate becomes larger than the service rate.

As expected, the average PDU transmission delay for BE traffic increases as the PDU arrival rate at the PS queue increases (Fig. 3). With a higher PDU arrival rate, since the PS queue requires more transmission bandwidth, the bandwidth allocated to the BE queue becomes smaller. Again, the simulation results closely follow the numerical results.

7.3.2 Performance of Queue-Aware Dynamic Bandwidth Allocation

The probability distributions for the allocated bandwidth to PS under different settings of the bandwidth adaptation thresholds (i.e., e) are shown in Fig. 4. As is evident, the distribution with smaller e results in higher variance than that with larger e. The higher variance indicates more fluctuations in the allocated bandwidth for PS.

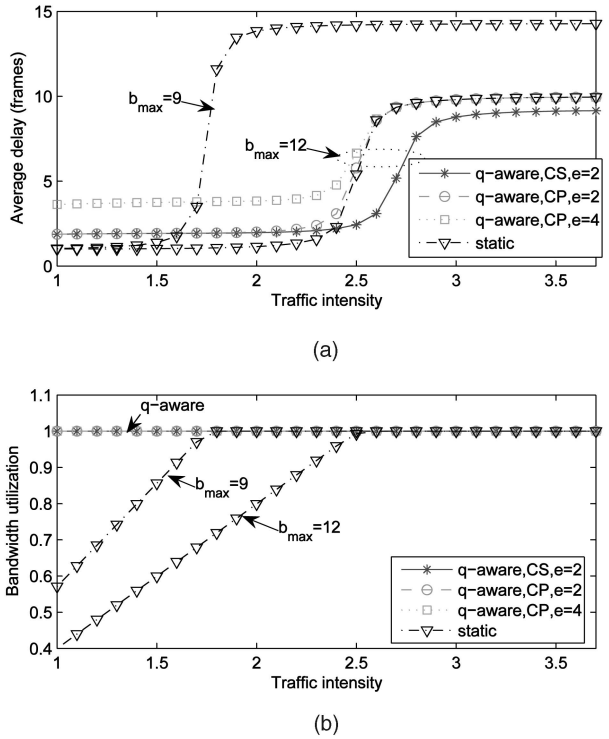


Fig. 5. Variations in (a) average delay and (b) bandwidth utilization under varying traffic intensity.

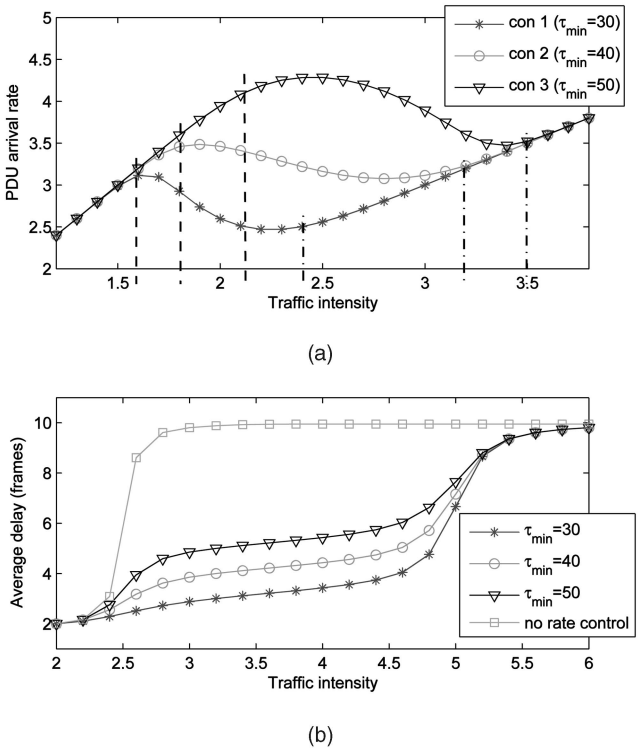


Fig. 6. Variations in (a) controlled PDU arrival rate for a PS connection and (b) average delay under different rate control threshold settings.

Fig. 5 a illustrates how the different threshold settings for dynamic bandwidth adaptation impacts the average delay for the PDUS in the PS queue. Specifically, larger e leads to higher average delay when the traffic intensity is low. The results for static bandwidth allocation are also shown for comparison.

With static allocation, at low traffic intensity delay is always one. The critical rate and the maximum average delay (i.e., average delay when the queue becomes unstable) depend on the amount of allocated bandwidth (i.e., b_{max}) to the SS. Interestingly, the proposed bandwidth allocation scheme with complete partitioning can maintain constant delay when the traffic intensity is low and the critical rates as well as the maximum average delay are equal to those for the case of static allocation when the traffic intensity is high. In case of complete sharing, the PS queue benefits from the off periods in the multistate on-off source and, therefore, the critical rate is higher and the maximum average delay is lower (e.g., 2.75 and 9 frames, respectively, in Fig. 5a). Also, we observe that the queue-aware allocation always achieves 100 percent utilization of the bandwidth (Fig. 5b).

Note that, while selecting the thresholds for dynamic bandwidth allocation, the value of e should not be too large so that the average delay can be kept small and, again, it should not be too small so that the high variability in the allocated bandwidth to the PS queue can be avoided (Fig. 5a and Fig. 4). The desired setting can be determined by using the analytical model.

7.3.3 Performance of the Queue-Aware Rate Control Scheme

Fig. 6a shows typical variations in the controlled PDU arrival rate for three different connections when the traffic intensity (per connection) increases. In this case, we set $\tau_{max} = 70$ and vary τ_{min}. With variation in traffic intensity, the controlled arrival rate decreases when the queue length becomes larger than the threshold τ_{min}. However, according to the modeling assumption, the controlled arrival rate cannot be reduced below the minimum guaranteed rate, which is half of the traffic intensity in this case. This explains the "ripple"-like behavior of the controlled arrival rate. Note that the threshold settings determine the values of the traffic intensity at which the slopes of the envelope of the controlled arrival rate change and the minimum guaranteed rate is achieved.

Typical variation in average delay for the PDUs in the PS queue under different rate control threshold settings for the PDUs in the PS queue is shown in Fig. 6b. Even though the average delay increases with increasing traffic intensity, due to rate control, the average delay does not approach maximum delay very rapidly as in the case without rate control. However, as the traffic intensity increases to a certain point (e.g., $\lambda = 5.5$ in Fig. 6b), there is no difference between any rate control threshold setting since the traffic sources reach their minimum guaranteed rates. Therefore, the average delay is close to the maximum delay, which indicates that the queue is full most of the time. Also, smaller τ_{min} results in lower delay since the PDU arrival rate is controlled earlier compared to the case with larger τ_{min}.

7.3.4 Transient Analysis

For transient analysis of the QoS performances of the adaptive bandwidth allocation and rate control schemes, we assume that the PS queue is empty at time zero (i.e., $\boldsymbol{\pi}_{tr}(0) = [1 \quad 0 \quad \cdots \quad 0]$). We vary the number of PS connections during different time periods (e.g., $N = 3, 4, 5, 6, 7, 5$ during time periods 1-40, 41-80, 81-120, 121-160, 161-200, and 201-240, respectively, in Fig. 7). We consider the complete partitioning case here with $b_{max} = 12$ and set the traffic intensity parameter to one (i.e., $\alpha = 1$).

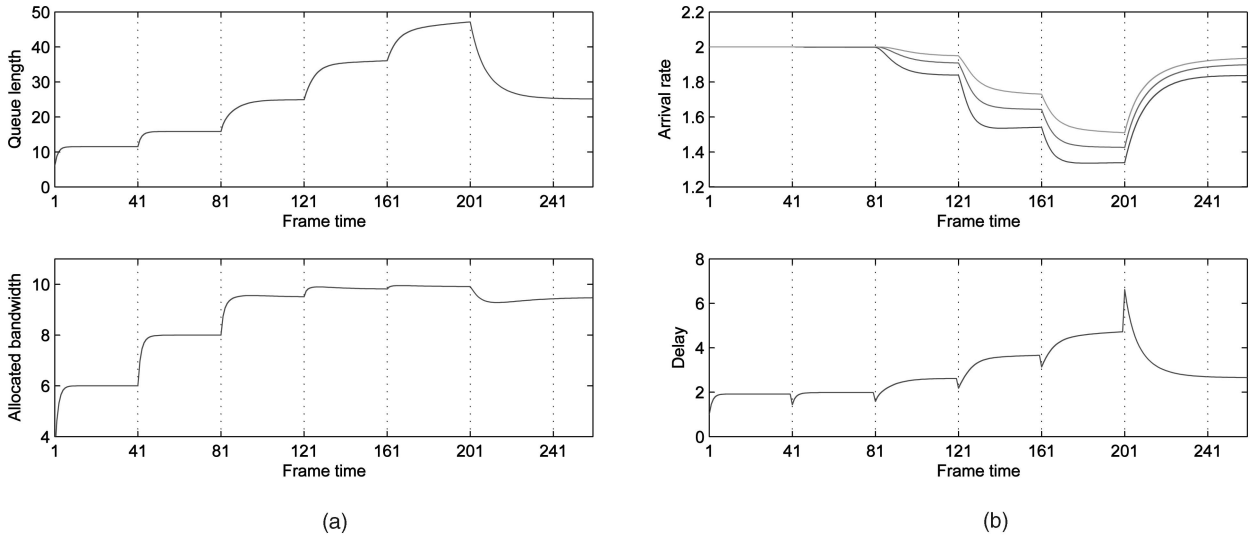


Fig. 7. (a) Queue-length and allocated bandwidth for PS queue and (b) controlled arrival rate and average delay obtained from transient analysis.

Typical variations in queue-length, amount of allocated bandwidth, controlled PDU arrival rate, and average delay with time are shown in Fig. 7. For controlled PDU arrival rate, we observe only the first three connections each of which has a different threshold settings (i.e., $\tau_{min} = 30, 40, 50$ and $\tau_{max} = 70$). For the other connections, we assume $\tau_{min} = 40$ and $\tau_{max} = 70$.

The PS queue length increases asymptotically (toward the average number of PDUs at steady state) with an increasing number of PS connections (Fig. 7a). With the queue-aware bandwidth allocation, when the number of PS connections becomes more than five (so that the sum of PDU arrival rates becomes larger than b_{max}), the allocated bandwidth reaches the maximum available bandwidth, at which point, the transmission rates for the connections are controlled (as shown in Fig. 7b). In this case, since different connections have different rate-control threshold settings, the arrival rate is controlled differently for each conection.

We observe from Fig. 7b that, when the number of connections is less than five, the average delay remains constant. However, when the queue becomes unstable, average delay is less than maximum average delay since the arrival rate for each of the connections is controlled. Note that, the discontinuities in the variation in average delay are due to the change in the number of PS connections which results in a sharp change in the PDU arrival rate into the queue. This causes transient variations in the amount of bandwidth allocation. Since the durations of these discontinuities are typically only a few frame intervals, the impact on overall performance would be negligible.

8 Conclusions

We have presented a queue-aware adaptive uplink bandwidth allocation and rate control mechanisms for polling service in IEEE 802.16 broadband wireless access networks. By utilizing the queue state information, the proposed mechanisms can maintain the packet-level QoS performances at the desired level. We have presented a comprehensive queuing analytical model to investigate the performances of the proposed schemes in both steady and transient states. While an exact model that captures the correlation of all traffic sources and the number of PDUs in the polling service queue has been formulated, an approximation model for the best-effort queue has been presented.

Performance evaluation of the proposed radio resource management model has been carried out extensively which reveals the interrelationships among the different performance measures. The correctness of the analytical model has been validated by simulations. The proposed analytical framework would be useful for 802.16-based radio access design and engineering.

The following provides a summary of the key results:

- When the PS queue is stable, the queue-aware uplink bandwidth allocation scheme can maintain the average delay at a constant level while maximizing the resource (i.e., radio bandwidth) utilization.
- While the maximum allocated bandwidth affects the critical rate and the maximum delay when the queue is unstable, the threshold settings affect the average delay when the queue is stable and also affects the distribution of bandwidth allocation. In particular, a larger interval between thresholds results in higher delay; however, the fluctuation in the allocated bandwidth becomes smaller.
- Since the setting of the rate-control thresholds (i.e., τ_{min} and τ_{max}) affects the QoS performances, rate control can be used to prioritize different PS connections through different threshold settings.
- The impacts of bandwidth allocation and rate-control parameter settings (for PS connections) on the performance of BE service can be quantified by the approximate queuing analytical model for BE service presented in this paper.

Acknowledgments

This work was supported by a scholarship from the TR*Labs*, Winnipeg, Canada, and in part by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

REFERENCES

[1] IEEE 802.16 Standard—Local and Metropolitan Area Networks —Part 16, IEEE Std 802.16a-2003, 2003.

[2] C. Eklund, R.B. Marks, K.L. Stanwood, and S. Wang, "IEEE Standard 802.16: A Technical Overview of the $WirelessMAN^{TM}$ Air Interface," *IEEE Comm. Magazine,* vol. 40, no. 6, pp. 98-107, June 2002.

[3] T.V.J. Ganesh Babu, T. Le-Ngoc, and J.F. Hayes, "Performance of a Priority-Based Dynamic Capacity Allocation Scheme for Wireless ATM Systems," *IEEE J. Selected Areas in Comm.,* vol. 19, no. 2, pp. 355-369, Feb. 2001.

[4] L. Muscariello, M. Meillia, M. Meo, M.A. Marsan, and R.L. Cigno, "An MMPP-Based Hierarchical Model of Internet Traffic," *Proc. IEEE Int'l Conf. Comm.,* vol. 4, pp. 2143-2147, June 2004.

[5] K. Wongthavarawat and A. Ganz, "Packet Scheduling for QoS Support in IEEE 802.16 Broadband Wireless Access Systems," *J. Comm. Systems,* vol. 16, pp. 81-96, 2003.

[6] K.K. Leung and A. Srivastava, "Dynamic Allocation of Downlink and Uplink Resource for Broadband Services in Fixed Wireless Networks," *IEEE J. Selected Areas in Comm.,* vol. 17, no. 5, pp. 990-1006, May 1999.

[7] G. Liu, W. Lang, W. Wu, Y. Ruan, X. Shen, and G. Zhu, "QoS-Guaranteed Call Admission Scheme for Broadband Multi-Services Mobile Wireless Networks," *Proc. IEEE Int'l Conf. Comm.,* vol. 1, pp. 454-459, June-July 2004.

[8] G. Li and H. Liu, "Dynamic Resource Allocation with Finite Buffer Constraint in Broadband OFDMA Networks," *Proc. IEEE Wireless Comm. and Networking Conf.,* vol. 2, pp. 1037-1042, Mar. 2004.

[9] M. Soleimanipor, W. Zhuang, and G.H. Freeman, "Optimal Resource Management in Wireless Multimedia Wideband CDMA Systems," *IEEE Trans. Mobile Computing,* vol. 1, no. 2, pp. 143-160, Apr.-June 2002.

[10] S. Baey, M. Dumas, and M.-C. Dumas, "QoS Tuning and Resource Sharing for UMTS WCDMA Multiservice Mobile," *IEEE Trans. Mobile Computing,* vol. 1, no. 3, pp. 221-235, July-Sept. 2002.

[11] J. Ye, J. Hou, and S. Papavassiliou, "A Comprehensive Resource Management Framework for Next Generation Wireless Networks," *IEEE Trans. Mobile Computing,* vol. 1, no. 4, pp. 249-264, Oct.-Dec. 2002.

[12] C.-T. Chou and K.G. Shin, "Analysis of Adaptive Bandwidth Allocation in Wireless Networks with Multilevel Degradable Quality of Service," *IEEE Trans. Mobile Computing,* vol. 3, no. 1, pp. 5-17, Jan.-Mar. 2004.

[13] Z. Liu, P. Nain, and D. Towlsey, "On Optimal Polling Policies," *Queueing Systems Theory and Applications,* vol. 11, no. 11, pp. 59-83, 1992.

[14] L. Kalampoukas, A. Varma, and K.K. Ramakrishnan, "Two-Way TCP Traffic over Rate Controlled Channels: Effects and Analysis," *IEEE/ACM Trans. Networking,* vol. 6, no. 6, pp. 729-743, Dec. 1998.

[15] H. Zhang, J. Cong, and O.W. Yang, "Rate Control over RED with Data Loss and Varying Delays," *Proc. IEEE GLOBECOM '03,* vol. 6, pp. 3035-3040, Dec. 2003.

[16] T. Inzerilli, "Design and Performance Modeling for Traffic Control in Wireless Links," *Proc. IEEE Int'l Conf. Comm.,* vol. 4, pp. 230-2311, June 2004.

[17] D.W. Dormuth and A.S. Alfa, "Two Finite-Difference Methods for Solving MAP(t)/PH(t)/1/K Queueing Models," *Queueing Systems,* vol. 27, pp. 55-78, 1997.

[18] M. Zorzi, "Packet Dropping Statistics of a Data-Link Protocol for Wireless Local Communications," *IEEE Trans. Vehicular Technology,* vol. 52, no. 1, pp. 71-79, Jan. 2003.

[19] Q. Liu, S. Zhou, and G.B. Giannakis, "Cross-Layer Combining of Queuing with Adaptive Modulation and Coding over Wireless Links," *Proc. IEEE Military Comm. Conf.,* vol. 1, pp. 717-722, Oct. 2003.

[20] P. Salvador, R. Valadas, and A. Pacheco, "Multiscale Fitting Procedure Using Markov Modulated Poisson Processes," *Telecomm. Systems,* vol. 23, pp. 123-148, 2003.

[21] M.F. Neuts, *Matrix Geometric Solutions in Stochastic Models—An Algorithmic Approach.* Baltimore: John Hopkins Univ. Press, 1981.

[22] I. Koffman and V. Roman, "Broadband Wireless Access Solutions Based on OFDM Access in IEEE 802.16," *IEEE Comm. Magazine,* vol. 40, no. 4, pp. 96-103, Apr. 2002.

Dusit Niyato (S'05) received the BSc degree in computer engineering from King Mongkut's Institute of Technology Ladkrabang (KMITL), Thailand, in 1999. In 2005, he received the MSc degree in electrical and computer engineering from the University of Manitoba, Canada. He is working toward the PhD degree in the Department of Electrical and Computer Engineering at the University of Manitoba. He is a researcher at TR*Labs*, Winnipeg, Canada. From 1999-2003, he worked as a researcher at the Embedded Systems Labs (ESL), Thailand. His main research interests are in the area of modeling, analysis, and optimization of protocols and architectures for broadband wireless networks. He is a student member of the IEEE.

Ekram Hossain (S'98-M'01-SM'06) received the BSc and MSc degrees, both in computer science and engineering, from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 1995 and 1997, respectively. He received the PhD degree in electrical engineering from the University of Victoria, Canada, in 2000. He is an associate professor (with tenture) in the Department of Electrical and Computer Engineering at the University of Manitoba, Winnipeg, Canada. He was a University of Victoria Fellow and also a recipient of the British Columbia Advanced Systems Institute (ASI) graduate student award. Dr. Hossain's research interests include radio link control and transport layer protocol design and cross-layer optimization issues for wireless networks, mobile computing, and distributed systems. He leads the Wireless Internet and Packet Radio Network Research Group in the Department of Electrical and Computer Engineering at University of Manitoba. Currently, he serves as an editor for the *IEEE Transactions on Wireless Communications*, the *IEEE/KICS Journal of Communications and Networks*, the *Wireless Communications and Mobile Computing Journal* (Wiley Interscience), and the *International Journal of Sensor Networks* (Inderscience Publishers). He served as one of the guest editors for the special issue of *IEEE Communications Magazine* on cross-layer protocol engineering for wireless mobile networks. He served as one of the guest editors for the special issue of the Wiley journal of *Wireless Communications and Mobile Computing* on radio link and transport protocol engineering for future-generation wireless mobile data networks. He also served as one of the guest editors for the special issue of the *IEEE Canadian Journal of Electrical and Computer Engineering* on advances in wireless communications and networking. Dr. Hossain served as a technical program cochair for the symposium on next generation mobile networks (NGMN '06) to be held in conjunction with the International Wireless Communications and Mobile Computing Conference (IWCMC '06), 3-6 July, 2006 in Vancouver, Canada. He served as a technical program committee member for the IEEE Globecom '06, ICC '06, ICC '05, WCNC '05, WCNC '04, Globecom '04, Globecom '03, and IFIP Networking '05. He was a recipient of the Lucent Technologies, Inc. research award for his contribution to the IEEE International Conference on Personal Wireless Communications (ICPWC) in 1997. He is a senior member of the IEEE.